Search Engine Technology and Digital Libraries

Libraries Need to Discover the Academic Internet

Norbert Lossau
Bielefeld University Library, Germany
<lossau@ub.uni-bielefeld.de>

This article is the revised and elaborated version of a presentation that was delivered at the invitation of the American Digital Library Federation (DLF) at their Spring Forum meeting in New Orleans (http://www.diglib.org/forums/Spring2004/springforum04abs.htm). It will be followed by "Search engine technology and digital libraries: Moving from theory to praxis" as a collaborative article from this author and Friedrich Summann, Head of IT at Bielefeld University Library.

With the development of the World Wide Web, the "information search" has grown to be a significant business sector of a global, competitive and commercial market. Powerful players have entered this market, such as commercial internet search engines, information portals, multinational publishers and online content integrators. Will Google, Yahoo or Microsoft be the only portals to global knowledge in 2010? If libraries do not want to become marginalized in a key area of their traditional services, they need to acknowledge the challenges that come with the globalisation of scholarly information, the existence and further growth of the academic internet [1].

How do libraries define academically relevant (online) content?

Libraries see themselves as central information providers for their clientele, at universities or research institutions. But how do they define academic content? Looking at the practice of today's digital library portals we get the impression that the internet is almost non-existent in the academic resource discovery environment. What we find are online library catalogues, electronic journals and (sometimes) e-books, which are mainly digitally converted print materials that have traditionally been the focus of library acquisition policies. Also databases have been well known for a long time. Content is generally delivered through well-established service channels by publishers, book-houses or subscription agencies.

The digitisation of publishing and the advent of the World Wide Web have resulted in the proliferation of a vast amount of content types and formats that include, but are not limited to, digitised collections, faculty and research groups' websites, conference web servers, preprint/e-print servers and, increasingly, institutional repositories and archives, as well as a wide range of learning objects and courses. If these resources are registered by a library at all, then they are in the form of separate lists of links or databases, but are not integrated into local digital library portals.

It is possible to identify reasons why libraries are hesitant to take action to change this situation. As a matter of principle libraries rank locally held collections and resources much more highly than remote resources, as the size of local collections has always been one indicator of the importance of a library. Libraries still see themselves as a place of collections rather than as an information "gateway". Other concerns of libraries are grounded in the fact that there is no guarantee that a remote host will maintain its resources in the long-term. Thus gateways to remote resources always have to face the potential problem of dead links. However, long-term accessibility is one of the basic values of libraries, and procedures need to be set up that will ensure that even remote repositories can be accessed in the long-term. Other reasons may include a natural resistance to the change of established acquisition procedures and workflows, as well as the complex combination of skills and competences that are required for "acquiring" remote resources, such as subject expertise, technical knowledge and traditional acquisition skills. This new type of "acquisition", introduced as a regular workflow within a library, would require some re-organisation of current structures with potential implications for costs and resources.

Are libraries aware of the incredible volume of academic content that is available on the web?

While libraries concentrate on the building of local digital library portals and simultaneous searches

across a selected number of licensed and free databases, do they see the incredible volume of academic content that is already available on the web? Although there are no reliable figures on the overall volume of web content there have been some studies that give estimations. Already published in 2001, a white paper from Michael Bergman on the "Deep Web" [2], highlights the dimensions we have to consider. Bergman talks about one billion individual documents in the "visible" [3] and nearly 550 billion documents on 200,000 web sites in the "deep" web. The exponential growth since 2001 can be read from the fact that in May 2004 Google gives the size of their index (i.e. visible web content) with more than 4,2 billion web pages (compared to 3,3 billion web pages in 2003).

For the research and teaching community the "invisible" web is of specific interest as it includes to a major proportion (high) quality content in free or licensed databases, primary data (e.g. meteorological, financial statistics, source data for bioresearch and so forth) or the huge and still increasing range of cultural and historical resources that are being digitised. There are, again, no reliable figures on the actual size of the academic web but the size of Scirus [4], the "science-specific search engine" of Elsevier might serve as a pure indicator for the volume dimension. In May 2004, 167 million science-specific Web pages have been indexed by Scirus, that are roughly 4% of the public Google-index (4,2 billion web pages). Although Scirus includes some "invisible" resources, the majority of the information has been crawled on web sites that are marked as "scientific" through their domain names [5]. Taking aside all the vagueness of these estimations one might apply the 4%-factor on the size of the invisible web in 2000 (i.e. 550 billion web pages) and receives the impressive figure of 22 billion web pages that include scientific content.

The vision...

How should libraries see the future of their information discovery services? Instead of a highly fragmented landscape that forces users to visit multiple, distributed servers, libraries will provide a search index, which forms a virtual resource of unprecedented comprehensiveness to any type and format of academically relevant content. Libraries liasing with other partners are contributing ultimately to an open, federated search index network that will offer an alternative to the monolithic structures of current commercial information.com indexes [6].

This unique resource will not form a minor segment within a commercial internet index, which lives from and is often heavily influenced by the advertisement industry, with their very specific rules about relevance and sustainability of information. Libraries will offer a long-term, reliable search service, which comprises high-quality content for the research and teaching communities.

Libraries are increasingly hesitant to support big, monolithic and centralised portal solutions equipped with an all-inclusive search interface which would only add another link to the local, customer-oriented information services. Future search services should be based on a collaboratively constructed, major shared data resource, but must come with a whole range of customisable search and browsing interfaces that can be seamlessly integrated into any local information portal, subject specific gateway or personal research and learning environment [7]. Libraries using the new search index must be able to select only those data segments that are of relevance specifically to their local or subject specific clientele, and search and browsing interfaces need to be customisable according to the local "look and feel" or discipline specific navigation mechanisms [8].

The new, academic search index should come with the ease of handling and the robustness and performance of Google-like services but with the relevance and proven ("certified") quality of content as it is traditionally made available through libraries.

Existing scholarly portal projects

Over the last few years the principle need to address the fragmentation of information resources has been recognised and has led to a number of initiatives at national and international levels. Very recently the German "Vascoda" portal [9] has been launched as a collaborative effort of German libraries, information providers and their international partners, jointly funded by the BMBF (Federal Ministry for Education and Research) and the DFG (Deutsche Forschungsgemeinschaft). The "Scholars Portal" initiative has been driven by the American Research Libraries (ARL) consortium and has produced a number of portal solutions on U.S. campuses [10]. The Resource Discovery Network (RDN) in the UK [11], virtual subject libraries and subject guides in Germany,

the European RENARDUS [12] project and the Internet SCOUT project (U.S.) [13] are only some examples of collaborative efforts to provide metadata based subject gateways to distributed content. Based on the OAI initiative, libraries and library service organisations are following the idea of "OAI Registries" as central points of access to worldwide distributed OAI repositories [14].

While undoubtedly successful in offering integrated access points, from the library point of view one gets the impression that there is still some development to be done in order to build real end-user services that find the full acceptance of researchers and students. In the era of popular internet full text search indexes these projects are focussing mainly on metadata by giving reference information about the resource (e.g. a certain server or database) rather than searching within the content sources (such as the full text itself). The records of all these portal databases, which usually describe intellectually selected content sources, can of course be used as a valuable starting point for the proposed discovery of the academic web. Where internet addresses are included in these records they can serve as starting URLs for web crawlers and other data aggregation tools that come with search engine technology [15]. It should be noted however that the major work for libraries building an academic web index will begin after the resource has been located, as a major proportion of the content in the academic web can not be aggregated by standard crawling mechanisms. This is why this part of the internet is called the "deep" or "invisible" web, and it comes, in particular in the academic environment, with an almost endless variety of data formats and technological implementations of databases and content servers.

The impact of internet search engines on libraries

It is a fact that with the advent of the World Wide Web, the information "search" has grown to be a significant business sector of a global, competitive and commercial market. Libraries are only one player within this market. Other stakeholders include, but are not limited to, publishers, online content integrators and commercial internet search engines ("information.coms").

In any market situation it is of paramount importance to take a close look at potential customers and their usage behaviour. For librarians this might sound obvious as it is their genuine perception that they consider implicitly the demands of users—or rather what they consider to be the demands of their users. But the new, competitive situation forces libraries to see things much more from the perspective of the user. First of all, this is an acknowledgement that, particularly at universities, libraries deal with a range of users with often different usage behaviours. It almost goes without saying that an undergraduate has other demands for information than a qualified researcher, and their usage behaviours can vary substantially. Young undergraduates will try much harder to transfer their general information seeking behaviour (using internet search engines) to the specific, academic environment, while established researchers have better accommodated the use of specific search tools. Before the WWW had been developed, this differentiation was, from the librarian's point of view, only relevant with respect to the level of training that various user groups required in order to use the library's resource discovery tools (printed catalogue, online catalogue, digital library portal). Today, with a whole range of general search engines available, users have the opportunity to use other catalogues (public or academic), and portals than those found in the library. Library users have been "empowered" by Google-like search engines to make their own choice about a search tool and to approach the world of information without any training. While librarians are mainly worried about the quality of information resources that are covered by mainstream search indexes, their users love these new tools and they would like to use them for any type of information search.

As a survey carried out at Bielefeld University in November 2002 revealed, students still make intensive use of the online library catalogue, but they would much prefer to access the catalogue through a "Google-like"-interface [16]. The simple reason why they still use the online catalogue is that, for this information type, they don't have an available alternative, as internet search engines usually don't cover the so called "deep" or "invisible" web. In any area where students think that they can find information, especially when they are looking for documents and full text, general search engines are even now much more popular than databases that have been made available through libraries. And it is only because of their level of experience and a certain habit that researchers still use databases, e-journals etc. that are not indexed by internet indexes. But a new generation of researchers will be coming in a few years time that will have grown up with popular internet search tools.

Just as users like the ease of phrasing and submitting a search query, they also like the flexible and responsive display of result sets. Superior performance and the size of internet search indexes are most impressive to them.

Challenges for search systems of current portal products

Bielefeld University Library has been leading the major digital library portal project for academic libraries in North Rhine-Westphalia, which has been running as a regular service since 2001 [17]. Based on this experience which resulted in a successful service only a few years ago, a new project group at the library looked into the recent development of the academic web and identified a list of limitations that come with current digital library portal systems:

Coverage of data formats, full text search

Most systems focus solely on the search of metadata (bibliographic fields, keywords, abstracts). The cross-search of full text has only recently been introduced and is often restricted to a very limited range of data formats (primarily "html" and "txt").

Coverage of content types

Current digital library systems integrate predominantly online library catalogues and databases with some full text repositories (e.g. e-journals). Freely available academic online content as described above is usually not covered by library portals. If they are selected at all they are mainly organised as html-link lists or specific databases (e.g. subject guides) that record reference metadata about web repositories.

Beyond online catalogues, databases and e-journals, researchers started to place their pre-prints or post-prints on the websites of faculties and research groups. Comprehensive web servers of scientific congresses include online presentations and papers, large international pre-print servers, often organised by the scientific community, store thousands and hundreds of thousands of documents, and the creation of e-learning objects is gaining increasing popularity.

And libraries? They add to the content that is available online. Today we have seen almost 15 years of digitisation activities, starting in the U.S. and spreading from there to other countries. Hundreds, if not thousands of digital document servers are available today, the majority of them as stand-alone systems. And activities at universities in building institutional repositories have only started. The long term goal is to store the research and e-learning output of each institution on self-controlled document servers. While the building of these repositories especially must be welcomed for strategic reasons (e.g. open access to research data, ensuring long term accessibility) the expected number of additional online hosts requires additional efforts on the search side.

Limited scalability / Information Retrieval performance

The majority of the portal systems rely on the metasearch (broadcast search) principle, i.e. a query is translated into the retrieval language of the target repositories (e.g. catalogues, databases) and sent out to selected repositories. The sequentially incoming responses are aggregated and presented in a joint result list.

The problems resulting from this search principle are well-known: due to the sequential response of the target repositories and in particular due to the dependence on the performance of these repositories we get—with an increasing number of target databases—limited scalability and decreasing performance.

Search comfort

All products follow the principles of traditional Boolean searching which has an impact on the ease of searching. While search engines, based on advanced linguistic analysis and semantic dictionaries, are increasingly integrating algorithms of approximate searching that allow a greater fault tolerance of search terms, traditional Boolean searches require some experience in finding suitable search

terms on the part of the user.

Presentation of result lists, relevance ranking of single hits

DL search systems rely on existing bibliographic fields such as author, title, and year of publication to sort their hit lists. The principles still follow, with some modifications, traditional print or card catalogues and are fairly inflexible. Search engines have introduced ranking algorithms that follow both static rules and dynamic live analysis of the result sets offering users a whole range of ranking criteria—one of the very popular features of internet search engines.

Shortcomings of "Information.com" internet indexes for the use of academic libraries

The discussion among libraries about their strategy for discovering the academic web has only started. Some institutions have decided to expose segments of their "invisible" content to Google (such as library catalogue records and institutional e-prints), a very pragmatic and cost-saving way to make quality content "visible" and at the same time searchable through this popular search service. While one can see the rationale in simply using existing internet services, libraries should not at this point omit to draft their own strategy and technical concepts for the academic internet.

In the first place, Google-like services are purely commercially driven. "A search engine's primary business is to obtain revenue through advertising" is the conclusion of Rita Vine in an article that has only recently been cited by Bonita Wilson as the editor of the *D-Lib Magazine* [18]. The ranking of search results follows several algorithms, one of which is the payment of search index providers through companies that want to boost their products.

Secondly, there is no guarantee of the long-term sustainability of an index, i.e. that the index of one week will include all resources of the last week. The dynamics and business concept of commercial indexes seems to be one of the top secrets of search index providers. The recent replacement of the former AllTheWeb-index delivered through Yahoo, with the own-brand Yahoo-index is one example that illustrates the uncertainty of indexes in the commercial search business. The new Yahoo-index is apparently smaller than AllTheWeb and it offers less functionality than AllTheWeb. Another example is the former "Google Directory", that featured a segment of the "Open Directory Project" and which has recently been replaced by Google with the shopping search engine "Froogle". Interestingly, this replacement took place initially only on the English Google-site, but not on others such as the German version. Let there be no misunderstanding: Commercial search indexes are within their rights and legitimacy to carry out whatever kind of business they feel is profitable. But libraries should not assume that these services will follow the rules of libraries rather than those of a profitable business.

Furthermore, commercial search indexes focus on content that can be automatically indexed. Manual conversion processes, time consuming analysis of data formats and access protocols is not in the interest of commercial services which need to increase their index sizes from week to week. Academic content, as described before, is often part of the invisible web and therefore not accessible to standard web robots per se. Institutions exposing their data to Google have to be aware that this can involve conversion work on the library site and that this conversion is mainly not conversion to a standard format like OAI with DC as supported by libraries.

Looking at the quality of content that is usually covered by information.coms, libraries have often expressed their concerns about the mixture of high quality academic and unproved mainstream content. Libraries are mainly missing the authoritative assessment of content resources as provided by current library search services, and also by those collaborative projects that have been described above.

Last but not least, the monolithic architecture of commercial indexes is very cost intensive, and experts speak of millions of USD to refresh one week's cycle of a major mainstream internet index [19]. There is no ONE library that could provide this kind of service on their own, thus libraries will need to rely on their traditional ability to collaborate in order to build a joint index network.

Apart from the obvious limitations that libraries should be aware of, there are also some myths around that should be named—and immediately dismissed. "Search engine technology can't handle

structured, high quality data" is as wrong as "Search engine technology comes only with simple (i.e. simplistic) search boxes". When statements like those described come up, people are simply mixing up some implementations with the actual potential of this technology.

Commercial internet search indexes vs. generic search engine technology

The majority of the objections amongst librarians against search engine technology are bound to the business concept of commercial indexes, not to the technology itself that lies behind the index. But why shouldn't libraries look beyond these current implementations and focus on the strengths of the generic technology?

If libraries look at large-scale, not specifically index-bound systems, there is essentially one technology company that is regularly listed as no. 1. The Norwegian company Fast Search & Transfer, an off-shoot of the Norwegian National University of Technology [20] is one of the market leaders in this sector and has been repeatedly awarded for technological innovation and excellence. The previous implementation of Fast as the search technology for internet indexes such as AllTheWeb and Altavista underline its ability to cope successfully with substantial volumes of data although Fast itself is explicitly branding their technology as enterprise search solution (ESP) to mark a wider application area than just the internet [21].

Bielefeld University Library, after an initial evaluation phase [22], has opted to use Fast as the technology to explore the benefits and usability for digital libraries. Rather than embarking on a new generic development of this technology from scratch there was a strategic decision to concentrate efforts on adding functionalities and modules to an existing product, such as data connectors (e.g. for OAI data, databases), new search and navigation interfaces or improved ranking and display features for academic content. Some of the positive aspects that came out of the evaluation of Fast software included comprehensive documentation, a whole suite of modules (e.g. linguistics, approximate search), powerful APIs and the support of standards (index works with XML queries). It is certainly not a turnkey solution but a system that proves to be very responsive to adaptations and customisation. Obviously the decision to use Fast was not an indefinite choice of a certain product but rather a pragmatic approach to collaborate with a partner that is innovation-oriented and which provides a robust and scalable core technology from day one. Alternatives, in particular open source developments like Nutch [23] and Jakarta Lucene [24], will be watched actively, and ways for collaboration will be explored.

Additional requirements for search services within the academic information environment

Beyond standard features of search engine technology, the discussion at Bielefeld University Library, in collaboration with colleagues from other libraries, resulted in a preliminary list of requirements which seem to be essential for implementations in the academic information environment:

- Indexing of qualified content resources only
  It has already been noted that existing databases and directories of the academic library community (such as virtual subject libraries and guides, regional and national portals) provide a profound base of intellectually selected content resources that can be indexed by those libraries that implement search engine technology. This approach adds an authoritative element to the envisaged index resource that is missing in most commercial internet indexes.

- Handling of data heterogeneity
  Search and navigation interfaces will need careful research and evaluation in order to confront potential users with a wide range of data formats and content types. Metadata, full text, images, multimedia objects and binary data are the more widely known examples of heterogeneous data, which can be aggregated in one virtual resource. The intelligent mark-up of certain content types for search, search refinement and navigation processes could be one way of dealing with data heterogeneity.

- Advanced navigation (browsing) functionality
  The term "search service" for the academic web always implies "navigation" of information. Established and partly implemented tools are scientific taxonomies, subject specific thesauri

and cross-concordances which support interdisciplinary research. Expected outcomes of the current semantic web research should also be evaluated for use within academic web searching.

- Flexible ranking and ordering schemes for result displays
  The result display of a new search service should be offered in a way that allows various views on a specific result set. Ranking and ordering schemes of search engine technology can principally follow pre-defined rules (such as sort by author, title, year, classification etc.) as well as ranking relevance generated on-the-fly (e.g. through live analysis of documents for semantics and syntactical structures).

- Automatic extraction of metadata
  Metadata have a high priority for the search and navigation of academic information. As a considerable proportion of resources (such as documents of scholars that are stored on faculty servers) have no descriptive information at all, it would be most helpful to develop document analysis tools that generate at least a minimum set of descriptive bibliographic information automatically [25].

Architecture of next generation search services

System modularity

Current search services are encapsulated and hard-coded within library management and digital library systems. This type of monolithic system architecture is not any longer state-of-the-art. Vendors of integrated library systems have partly responded to this development and offer already separate local and central (i.e. shared cataloguing) modules. New requirements for libraries have resulted in the set-up of new systems such as digital library systems, digital collection and e-print servers. And the range of systems extends substantially if we take a wider look within a university, beyond the actual library environment. The increase of systems alongside with the increased demand on financial and staff resources to maintain these systems have led to discussions within libraries and on a campus wide level in order to find out a) how theses systems interact with each other and b) investigate potential duplication or even multiplication of services implemented in different systems. A widely known example is the administration of users at a university and the discussion about "single sign-on" mechanisms across systems. Also new areas like e-Learning have revealed new demands to re-use existing services in learning management environments.

System modularity can also work for digital library and scholarly portals. A portal comprises the gateway to distributed, heterogeneous resources plus additional services. In the Digital Library North Rhine-Westphalia one can find for example a "metasearch" as gateway to more than 60 databases and catalogues and as additional services e.g. the check of local availability of resources and the set-up of personal search profiles. On the medium-term web services could be used within existing digital library or other information portals to replace current search techniques by state-of-the-art search engine technology.

Interoperable services

The technology of web-services provides true interoperability between services and system modules from different system environments. They work across all system platforms ("platform-independence"), are easy to implement and open the opportunity to realise local integrations of external systems. At Bielefeld University the library in collaboration with the IT-department of the central administration worked successfully on the first implementation of web-services for central user-directories and the online library catalogue. The new technical modularity enables and stimulates a much more holistic view on the university's architecture and is a promising trend to overcome all the smaller and bigger silo-solutions that we currently see at a university.

Bielefeld UL's concept for an open, federated academic internet index network, as described in the next chapter, takes up these developments of modular architectures. It proposes the implementation of web services to support actively the interoperability of distributed search indexes and seamless integration into local, regional or subject based information infrastructures.

Libraries to build an open, federated academic internet index network?

The continual exponential growth in the volume of online web content as described above makes it unrealistic to believe that one library can build one big, all-inclusive academic web index. Even to provide a substantial part, such as indexing the academic online content of one country, would mean a major challenge to one institution. Thus, collaboration is required among libraries -- and libraries have an excellent tradition in collaborating, including in the international context. How could the collaboration work in practice?

Is the technology for a federated web index available?

Commercial internet indexes usually have a monolithic architecture [26]. The technology itself allows already federated indexes within specific scenarios. Fast's software suite is familiar with the building of a virtual master-index that is comprised of distributed child-indexes. This requires the use of Fast technology on all partner sides. Other business implementations of Fast software also integrate result sets of external internet search indexes. Despite some functionalities that support federated indexing, this area will require further research.

The current Grid-research initiatives, that address distributed, large-scale computing in a wider context, could provide valuable technology for the building of distributed data and access networks. Libraries will need to watch closely these developments and be open for collaborations.

Potential roles for partners in the academic web index network

There are different roles for partners, that can be adopted according to local financial and staff resources or individual policies etc.:

1. The "user": Seamless integration of search and navigation interfaces into local environments
The easiest way to participate is to simply use the new index resource and make it popular within the local academic communities. The "local use"—concept of internet indexes has been made popular by Google and other commercial index providers through the "search box"—feature that can be freely incorporated into the local information infrastructure. Submitted searches go directly to the index and hits are returned as standard result sets of the provider with standard display and ordering rules.

The new academic index network will offer the same functionality. It doesn't require any software investment [27] and only minimum staff resources [28].

Any customisation to these interfaces such as the restriction to specific content resources in order to meet specific local needs (e.g. only certain subject areas) or the implementation of specific search or browsing features (such as the use of a specific classification) will have to be realised on the server side by one of the index network partners that host the respective proportion of the search index. In the case of customisation work (i.e. programming) this will have to be refunded by the requesting local institution.

2. The "content provider": Exposing online library collections and other academic content to the "index creators"
Beyond using the new search service all academic content providers (and libraries in particular) are encouraged to make their current "invisible" content accessible to the index creating partner libraries.

During the last year of exploring search engine technology and setting up demonstrators Bielefeld UL has been supported by a number of libraries in Germany, the UK and the U.S. through the free exposure of content. Data have been loaded, (pre-)processed and indexed from more than 10 institutions including the Mathematical Faculty at Bielefeld University, Göttingen State and University Library, The National Library for Technology (TIB, Hanover), Bielefeld University Library, Oxford University Library Services, Michigan and Cornell University Libraries, Springer publisher and the "Zentralblatt für Mathematik" database. A range of different content types with a wide range of formats included digitised collections, pre-print servers, electronic journals,

institutional repositories, online library catalogues and databases. One method of loading data was through OAI harvesters which revealed a number of problems as discussed before. Essential for all partners was that they only exposed their data to the aggregating tool (like web crawler, ftp, metadata harvester), i.e. all collections resided on-site within the control of each partner. This has proved to be fundamental both for commercial partners and for public libraries and institutions. Commercial partners in particular recognised that this new search service of a library can create additional access points to their content without affecting existing licensing agreements at all. The new academic search index, analogue to commercial internet indexes, simply offers a new, comfortable and high quality search and navigation platform to distributed content, while the use of content itself still happens on the content provider side.

A second important aspect was that the identity of each partner and collection has been clearly marked within the result sets by adding a clause to each hit that says: "provided by institution XYZ".

3. The "index creator and host": Building local, regional or subject specific search indexes and sharing them within the federated network

Libraries that are interested in building up their own expertise with search engine technology can create and host their own search index. Those institutions will need to provide a technological infrastructure comprising hardware [29], search engine technology and staff resources to handle the data workflow. Bielefeld UL has identified five general working steps that need to be carried out once a content resource has been selected: Data loading, pre-processing, processing, indexing and access and navigation. Staff should be familiar with technologies like harvesting tools (e.g. for OAI), database connectors and protocols (like Z39.50, SRW, generic db-connectors), Perl and XSLT (pre-processing), Python (processing), specific search engine technology (indexing) and finally PHP (access and navigation).

There are numerous selection criteria for libraries when they start to think about building "their" index and obviously the local demands will have top priority for the selection process. If libraries think beyond their local index and wish to participate in a shared effort of the international library community, an action plan will be needed in order to avoid unnecessary duplication of work [30]. Libraries could work according to geographical criteria as well as subject expertise and pragmatic approaches can also address specific content resource types such as OAI repositories, institutional servers, digital collections etc.

Potential partners in this group include all projects and initiatives that have been described above as "existing scholarly portal projects" such as providers of subject guides, hubs and other information gateways. Organisation and communication will be paramount in order for this initiative to be successful.

4. The "technology support partners": Developing and sharing tools that help to build high quality search services

There are many areas where work-sharing on the technical side would help to create synergies, save resources and speed-up the process of building a high quality virtual resource of academic web content.

Major work needs to be done in order to give strong support for automatic data loading and (pre-)processing processes. An important contribution needs to come from the OAI community by working on a much more consistent use of these promising standards. If a standard like OAI already causes problems one can imagine the dimension of this work when it comes to the huge amount of resources that are not OAI-compliant. Bielefeld UL, based on its experience with integrating data resources into the "Metasearch" of the Digital Library portal, has already developed some scripts and processing stages that were required to integrate only 15 different repositories into the current search engine implementation.

On the search and navigation interface side the integration of taxonomies, classifications etc. is another key area that could be addressed in a collaborative approach.

5. Non-library partners

Some commercial content providers such as Elsevier and Thomson have already replaced their search technology and created new indexes [31]. It seems to be reasonable to integrate existing indexes into the community-driven web index network if those index providers are prepared to

agree on the terms and conditions which need to be defined by the library community.

Conclusion and outlook

Will Google, Yahoo or Microsoft be the only portals to global knowledge in 2010? This paper advocates a concerted initiative of the library community to pick up state-of-the-art search technology and build reliable, high quality search services for the research and teaching community. This effort is not intended as competition to other commercial services, but it represents a natural continuation of traditional library services in a globalised academic information environment. An academic internet index network, driven by the library community as sketched out before, can best meet the specific requirements of the scholarly clientele by providing comfortable, reliable and integrated access to high quality content. But in order to realise this new service libraries are required to look beyond their current information infrastructure to learn from mainstream internet index providers who have become so popular through innovative technology and a dedicated end-user driven approach.

How can libraries proceed?

The library community needs to acknowledge the relevance of a new action plan in order to improve current search services. The impression is that many libraries "somehow" see the need but it's still unclear for them how to address the problem. Current pragmatic approaches to make academic content available to commercial internet indexes should be seen only as a first step on the way to a new service that is driven by the libraries themselves.

Bielefeld UL sees the further development of search services as mission-critical to their role as central information provider for the university. The library is currently working on the first public version of BASE (Bielefeld Academic Search Engine) and details about the practical experiences with search engine technology will be discussed in part II of this article [32]. Bielefeld UL has been successful in carrying out a proof-of-concept by implementing search engine technology and working with real-life resources. The first public demonstrator of their search engine technology implementation (BASE) is now available: the demonstrator for a "Digital Collections" search service [33]. A second demonstrator that focuses on mathematical literature will be available shortly. The conclusions and experiences of this work have been included in a current joint proposal from Bielefeld and the Regional Service Centre for Academic Libraries in North Rhine-Westphalia to the Deutsche Forschungsgemeinschaft (DFG). Both the DFG and the Federal Ministry of Education and Research have decided to fund a collaborative initiative called "Distributed Document Server" (VDS) within the German Digital Library, "Vascoda". Three major lines of action have been defined for the VDS project: authentication, metadata registry and seamless navigation with search engine technology. Bielefeld UL welcomes any collaboration and will be very happy to share their part of an online index nucleus with other library partners in order to build the suggested network.

Beyond Germany other libraries and institutions such as Oxford University Library Services and the National Library of Norway have expressed their explicit interest in the new technology. A number of partner institutions within the American Digital Library Federation have pledged their support for a joint, concerted initiative. Any other library, institution or content provider is encouraged to support the forthcoming activities. While Bielefeld University Library has been very successful in collaborating with the Fast company and will continue to expand this partnership, other institutions might select alternative systems. In fact, the only thing of major importance will be the openness of each of these systems to allow their interoperability for the suggested federated index network. If the concept itself can be realised, the choice for the actual system will, in any case, depend on a technical benchmarking process that must also include existing and forthcoming mainstream internet search services, as this is ultimately what is critical for acceptance by end-users.

It is expected that other libraries will start to create their own local search engine infrastructures in order to build further indexes. An (informal) network or forum will be formed where knowledge, content and tools can be shared. The need for trans-national action has never been so obvious as it is now, and it is hoped that funding organisations in many countries will acknowledge the dimension of this undertaking and give the support libraries need to fulfil their mission: to discover the rich wealth of the academic internet for the benefit of the international research and teaching community.

Acknowledgement

Acknowledgement for this article goes to the technical search engine working group at Bielefeld University Library (Friedrich Summann, Bernd Fehling, Carola Kummert and Sebastian Wolf) who have done exceptionally well over more than one year in exploring and implementing search engine technology and provided a lot of detailed insights.

Notes

[1] Very recently Stephen Arnold has addressed the threat that libraries could become marginalized in his paper "Information boundaries and libraries", February 2004 (http://www.arnoldit.com/articles/elib_Feb2004_final.doc).

[2] Michael K. Bergman (on behalf of the BrightPlanet Corporation) "The deep web: Surfacing hidden value". In: *The Journal of Electronic Publishing*. Michigan University Press. July 2001 (http://www.press.umich.edu/jep/07-01/bergman.html).

[3] Bergman uses the term "surface" web.

[4] See <http://www.scirus.com>.

[5] The majority of the deep web content comes from Medline and Elsevier's Science Direct database. Web sites that have been crawled include domains like ".edu": 58.5 million pages, ".ac.uk": 6.8 million pages (http://www.scirus.com/srsapp/aboutus/#range).

[6] Partners can in principle include library service centres and also publishers or other content integrating commercial services.

[7] Web services seem to be a promising technology that allow the modular integration of remote services into a local environment.

[8] Using navigation tools like thesauri, cross-concordances or classifications.

[9] See <http://www.vascoda.de>.

[10] See <http://www.arl.org/access/scholarsportal/>.

[11] See <http://www.rdn.ac.uk/>.

[12] See <http://www.renardus.org>.

[13] See <http://scout.wisc.edu/>.

[14] Three initiatives were present at the DLF Spring Forum 2004 by the University of Michigan Libraries (http://www.diglib.org/forums/Spring2004/hagedorn0404.htm), the University of Illinois, Urbana-Champaign (http://www.diglib.org/forums/Spring2004/spring04bios.htm#habing) and OCLC (http://www.diglib.org/forums/Spring2004/young0404.htm).

[15] The Fast software, for example, also includes a file traverser and some generic database connectors to load data from heterogeneous content sources.

[16] 625 undergraduates and graduates returned the questionnaire within a week.

[17] See <http://www.ub.uni-bielefeld.de/portal/english/Metasuche>. The development of the Digital Library North Rhine-Westphalia was coordinated by the former director of Bielefeld University Library, Dr. Karl Wilhelm Neubauer and funded through a major grant of the State Ministry for Science and Research. The project began in 1998 and officially ended in 2001 with the regular service coming into operation at the regional Service Centre for Academic Libraries in Cologne (HBZ). The portal technology (IPS system) has been developed by IHS according to the specifications made by the project's development group.

[18] Bonita Wilson. "Using the Internet for Searching," *D-Lib Magazine*, March 2004, Volume 10

Number 3, <doi:10.1045/march2004-editorial>.

[19] Stephen Arnold provided a figure of "$24 million per year to index one billion content sources". In: "The future of search", *Proceedings of the 26th Online Information conference 2002*. p 51.

[20] Fast Search & Transfer is based in Oslo, Norway with branches in some European countries, the U.S. and Japan (http://www.fastsearch.com). After selling their internet assets (like AllTheWeb) to Overture in April 2003 Fast has been focussing on innovative developments of their search technology through a major company research lab in Norway and external research collaborations with the Universities of Munich (Germany), Cornell, Penn. State (U.S.) and Trondheim/Tromsoe (Norway).

[21] Customers come from multinational enterprises, e-Government, e-Health etc.

[22] February 2002 to August 2002, Convera, Google, Mnogo, Lucene (2003).

[23] See <http://www.nutch.org>.

[24] See <http://jakarta.apache.org/lucene/docs/index.html>.

[25] Bielefeld UL is collaborating with the external research laboratory of Fast in Munich, Germany.

[26] Although the index data itself can be, for performance reasons, distributed on large clusters of computers.

[27] There is no local client software that needs to be licensed and installed.

[28] E.g., for the integration of search templates within local web sites.

[29] In general low-cost hardware such as Linux-PCs with appropriate mass-storage attached. Further details can be obtained from the technical project group at Bielefeld UL.

[30] Obviously a principle that applies to many other areas in the library world.

[31] Elsevier is using the Fast search software for some of their "integrated search" products (www.scirus.com and www.scopus.com), Thomson's Web of Knowledge is based on WebFeat Prism technology, <http://isiwebofknowledge.com/technology.html>.

[32] The article "Moving from theory to praxis" is scheduled to appear in an issue of *D-Lib-Magazine* later this year. A PowerPoint presentation that describes some of the details can be viewed on the website of the Digital Library Federation <http://www.diglib.org/forums/Spring2004/summann0404.htm>.

[33] See <http://base.ub.uni-bielefeld.de/>. Core technology for the search index is Fast Data Search; additional data processing scripts and user-interfaces have been developed at Bielefeld UL with other programming languages and open source technology.

---

doi:10.1045/june2004-lossau